The Cramér–Rao Bound for Sparse Estimation

Zvika Ben-Haim, *Student Member, IEEE,* and Yonina C. Eldar, *Senior Member, IEEE*

***Abstract*— The goal of this paper is to characterize the best achievable performance for the problem of estimating an unknown parameter having a sparse representation. Specifically, we consider the setting in which a sparsely representable deterministic parameter vector is to be estimated from measurements corrupted by Gaussian noise, and derive a lower bound on the mean-squared error (MSE) achievable in this setting. To this end, an appropriate definition of bias in the sparse setting is developed, and the constrained Cramér–Rao bound (CRB) is obtained. This bound is shown to equal the CRB of an estimator with knowledge of the support set, for almost all feasible parameter values. Consequently, in the unbiased case, our bound is identical to the MSE of the oracle estimator. Combined with the fact that the CRB is achieved at high signal-to-noise ratios by the maximum likelihood technique, our result provides a new interpretation for the common practice of using the oracle estimator as a gold standard against which practical approaches are compared.**

EDICS Topics: SSP-PARE, SSP-PERF.

Index terms: Constrained estimation, Cramér–Rao bound, sparse estimation.

I. Introduction

The problem of estimating a sparse unknown parameter vector from noisy measurements has been analyzed intensively in the past few years [1]–[4], and has already given rise to numerous successful signal processing algorithms [5]–[9]. In this paper, we consider the setting in which noisy measurements of a deterministic vector $\boldsymbol{x}_0$ are available. It is assumed that $\boldsymbol{x}_0$ has a sparse representation $\boldsymbol{x}_0 = \boldsymbol{D}\boldsymbol{\alpha}_0$, where $\boldsymbol{D}$ is a given dictionary and most of the entries of $\boldsymbol{\alpha}_0$ equal zero. Thus, only a small number of "atoms," or columns of $\boldsymbol{D}$, are required to represent $\boldsymbol{x}_0$. The challenges confronting an estimation technique are to recover either $\boldsymbol{x}_0$ itself or its sparse representation $\boldsymbol{\alpha}_0$. Several practical approaches turn out to be surprisingly successful in this task. Such approaches include the Dantzig selector (DS) [4] and basis pursuit denoising (BPDN), which is also referred to as the Lasso [1], [2], [10].

A standard measure of estimator performance is the mean-squared error (MSE). Several recent papers analyzed the MSE obtained by methods such as the DS and BPDN [4], [11]. To determine the quality of estimation approaches, it is of interest to compare their achievements with theoretical performance limits: if existing methods approach the performance bound, then they are nearly optimal and further improvements in the current setting are impossible. This motivates the development of lower bounds on the MSE of estimators in the sparse setting.

Since the parameter to be estimated is deterministic, the MSE is in general a function of the parameter value. While there are lower bounds on the worst-case achievable MSE among all possible parameter values [12, §7.4], the actual performance for a specific value, or even for most values, might be substantially lower. Our goal is therefore to characterize the minimum MSE obtainable for each particular parameter vector. A standard method of achieving this objective is the Cramér–Rao bound (CRB) [13], [14].

The fact that $\boldsymbol{x}_0$ has a sparse representation is of central importance for estimator design. Indeed, many sparse estimation settings are underdetermined, meaning that without the assumption of sparsity, it is impossible to identify the correct parameter from its measurements, even without noise. In this paper, we treat the sparsity assumption as a deterministic prior constraint on the parameter. Specifically, we assume that $\boldsymbol{x}_0 \in \mathcal{S}$, where $\mathcal{S}$ is the set of all parameter vectors which can be represented by no more than s atoms, for a given integer s.

Our results are inspired by the well-studied theory of the constrained CRB [15]–[18]. This theory is based on the assumption that the constraint set can be defined using the system of equations $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$, $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$, where $\boldsymbol{f}$ and $\boldsymbol{g}$ are continuously differentiable functions. The resulting bound depends on the derivatives of the function $\boldsymbol{f}$. However, sparsity constraints cannot be written in this form. This necessitates the development of a bound suitable for non-smooth constraint sets [19]. In obtaining this modified bound, we also provide new insight into the meaning of the general constrained CRB. In particular, we show that the fact that the constrained CRB is lower than the unconstrained bound results from an expansion of the class of estimators under consideration.

With the aforementioned theoretical tools at hand, we obtain lower bounds on the MSE in a variety of sparse estimation problems. Our bound limits the MSE achievable by any estimator having a pre-specified bias function, for each parameter value. Particular emphasis is given to the unbiased case; the reason for this preference is twofold: First, when the signal-to-noise ratio (SNR) is high, biased estimation is suboptimal. Second, for high SNR values, the unbiased CRB is achieved by the maximum likelihood (ML) estimator.

While the obtained bounds differ depending on the exact problem definition, in general terms and for unbiased estimation the bounds can be described as follows. For parameters having maximal support, i.e., parameters whose representation requires the maximum allowed number s of atoms, the lower bound equals the MSE of the "oracle estimator" which knows the locations (but not the values) of the nonzero representation elements. On the other hand, for parameters which do not have maximal support (a set which has Lebesgue measure zero in $\mathcal{S}$), our lower bound is identical to the CRB for an unconstrained problem, which is substantially higher than the oracle MSE.

Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel. Phone: +972-4-8294700, fax: +972-4-8295757, E-mail: {zvikabh@tx, yonina@ee}.technion.ac.il. This work was supported in part by the Israel Science Foundation under Grant no. 1081/07 and by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM++ (contract no. 216715).

The correspondence between the CRB and the MSE of the oracle estimator (for all but a zero-measure subset of the feasible parameter set $\mathcal{S}$) is of practical interest since, unlike the oracle estimator, the CRB is achieved by the ML estimator at high SNR. Our bound can thus be viewed as an alternative justification for the common use of the oracle estimator as a baseline against which practical algorithms are compared. This gives further merit to recent results, which demonstrate that BPDN and the DS both achieve near-oracle performance [4], [11]. However, the existence of parameters for which the bound is much higher indicates that oracular performance cannot be attained for *all* parameter values, at least using unbiased techniques. Indeed, as we will show, in many sparse estimation scenarios, one cannot construct *any* estimator which is unbiased for all sparsely representable parameters.

Our contribution is related to, but distinct from, the work of Babadi et al. [20], in which the CRB of the oracle estimator was derived (and shown to equal the aforementioned oracle MSE). Our goal in this work is to obtain a lower bound on the performance of estimators which are not endowed with oracular knowledge; consequently, as explained above, for some parameter values the obtained CRB will be higher than the oracle MSE. It was further shown in [20] that when the measurements consist of Gaussian random mixtures of the parameter vector, there exists an estimator which achieves the oracle CRB at high SNR; this is shown to hold on average over realizations of the measurement mixtures. The present contribution strengthens this result by showing that for any given (deterministic) well-behaved measurement setup, there exists a technique (namely, the ML estimator) achieving the CRB at high SNR. Thus, convergence to the CRB is guaranteed for all measurement settings, and not merely when averaging over an ensemble of such settings.

The rest of this paper is organized as follows. In Section II, we review the sparse setting as a constrained estimation problem. Section III defines a generalization of sparsity constraints, which we refer to as locally balanced constraint sets; the CRB is then derived in this general setting. In Section IV, our general results are applied back to some specific sparse estimation problems. In Section V, the CRB is compared to the empirical performance of estimators of sparse vectors. Our conclusions are summarized in Section VI.

Throughout the paper, boldface lowercase letters $\boldsymbol{v}$ denote vectors while boldface uppercase letters $\boldsymbol{M}$ denote matrices. Given a vector function $\boldsymbol{f}:\mathbb{R}^n\to\mathbb{R}^k$, we denote by $\partial\boldsymbol{f}/\partial\boldsymbol{x}$ the $k\times n$ matrix whose ijth element is $\partial f_i/\partial x_j$. The support of a vector, denoted $\operatorname{supp}(\boldsymbol{v})$, is the set of indices of the nonzero entries in $\boldsymbol{v}$. The Euclidean norm of a vector $\boldsymbol{v}$ is denoted $\|\boldsymbol{v}\|_2$, and the number of nonzero entries in $\boldsymbol{v}$ is $\|\boldsymbol{v}\|_0$. Finally, the symbols $\mathcal{R}(\boldsymbol{M})$, $\mathcal{N}(\boldsymbol{M})$, and $\boldsymbol{M}^\dagger$ refer, respectively, to the column space, null space, and Moore–Penrose pseudoinverse of the matrix $\boldsymbol{M}$.

II. Sparse Estimation Problems

In this section, we describe several estimation problems whose common theme is that the unknown parameter has a sparse representation with respect to a known dictionary. We then review some standard techniques used to recover the unknown parameter in these problems. In Section V we will compare these methods with the performance bounds we develop.

A. The Sparse Setting

Suppose we observe a measurement vector $\boldsymbol{y}\in\mathbb{R}^m$, given by

$$\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}_0+\boldsymbol{w} \tag{1}$$

where $\boldsymbol{x}_0\in\mathbb{R}^n$ is an unknown deterministic signal, $\boldsymbol{w}$ is independent, identically distributed (IID) Gaussian noise with zero mean and variance σ^2, and $\boldsymbol{A}$ is a known $m\times n$ matrix. We assume the prior knowledge that there exists a sparse representation of $\boldsymbol{x}_0$, or, more precisely, that

$$\boldsymbol{x}_0\in\mathcal{S}\triangleq\left\{\boldsymbol{x}\in\mathbb{R}^n:\boldsymbol{x}=\boldsymbol{D}\boldsymbol{\alpha},\|\boldsymbol{\alpha}\|_0\le s\right\}. \tag{2}$$

In other words, the set $\mathcal{S}$ describes signals $\boldsymbol{x}$ which can be formed from a linear combination of no more than s columns, or atoms, from $\boldsymbol{D}$. The dictionary $\boldsymbol{D}$ is an $n\times p$ matrix with $n\le p$, and we assume that $s<p$, so that only a subset of the atoms in $\boldsymbol{D}$ can be used to represent any signal in $\mathcal{S}$. We further assume that $\boldsymbol{D}$ and s are known.

Quite a few important signal recovery applications can be formulated using the setting described above. For example, if $\boldsymbol{A}=\boldsymbol{I}$, then $\boldsymbol{y}$ consists of noisy observations of $\boldsymbol{x}_0$, and recovering $\boldsymbol{x}_0$ is a denoising problem [5], [6]. If $\boldsymbol{A}$ corresponds to a blurring kernel, we obtain a deblurring problem [7]. In both cases, the matrix $\boldsymbol{A}$ is square and invertible. Interpolation and inpainting can likewise be formulated as (1), but in those cases $\boldsymbol{A}$ is an underdetermined matrix, i.e., we have $m<n$ [9]. For all of these estimation scenarios, our goal is to obtain an estimate $\widehat{\boldsymbol{x}}$ whose MSE is as low as possible, where the MSE is defined as

$$\mathrm{MSE}\triangleq E\left\{\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\|_2^2\right\}. \tag{3}$$

Note that $\boldsymbol{x}_0$ is deterministic, so that the expectation in (3) (and throughout the paper) is taken over the noise $\boldsymbol{w}$ but not over $\boldsymbol{x}_0$. Thus, the MSE is in general a function of $\boldsymbol{x}_0$.

In the above settings, the goal is to estimate the unknown signal $\boldsymbol{x}_0$. However, it may also be of interest to recover the coefficient vector $\boldsymbol{\alpha}_0$ for which $\boldsymbol{x}_0=\boldsymbol{D}\boldsymbol{\alpha}_0$, e.g., for the purpose of model selection [1], [4]. In this case, the goal is to construct an estimator $\widehat{\boldsymbol{\alpha}}$ whose MSE $E\{\|\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0\|_2^2\}$ is as low as possible. Unless $\boldsymbol{D}$ is unitary, estimating $\boldsymbol{\alpha}_0$ is not equivalent to estimating $\boldsymbol{x}_0$. Note, however, that when estimating $\boldsymbol{\alpha}_0$, the matrices $\boldsymbol{A}$ and $\boldsymbol{D}$ can be combined to obtain the equivalent problem

$$\boldsymbol{y}=\boldsymbol{H}\boldsymbol{\alpha}_0+\boldsymbol{w} \tag{4}$$

where $\boldsymbol{H}\triangleq\boldsymbol{A}\boldsymbol{D}$ is an $m\times p$ matrix and

$$\boldsymbol{\alpha}_0\in\mathcal{T}=\{\boldsymbol{\alpha}\in\mathbb{R}^p:\|\boldsymbol{\alpha}\|_0\le s\}. \tag{5}$$

Therefore, this problem can also be seen as a special case of (1) and (2). Nevertheless, it will occasionally be convenient to refer specifically to the problem of estimating $\boldsymbol{\alpha}_0$ from (4).

Signal estimation problems differ in the properties of the dictionary $\boldsymbol{D}$ and measurement matrix $\boldsymbol{A}$. In particular, problems of a very different nature arise depending on whether the dictionary is a basis or an overcomplete frame. For example, many approaches to denoising yield simple shrinkage techniques when $\boldsymbol{D}$ is a basis, but deteriorate to NP-hard optimization problems when $\boldsymbol{D}$ is overcomplete [21].

A final technical comment is in order. If the matrix $\boldsymbol{H}$ in (4) does not have full column rank, then there may exist different feasible parameters $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ such that $\boldsymbol{H}\boldsymbol{\alpha}_1 = \boldsymbol{H}\boldsymbol{\alpha}_2$. In this case, the probability distribution of $\boldsymbol{y}$ will be identical for these two parameter vectors, and the estimation problem is said to be unidentifiable [22, §1.5.2]. A necessary and sufficient condition for identifiability is

$$\operatorname{spark}(\boldsymbol{H}) > 2s \tag{6}$$

where $\operatorname{spark}(\boldsymbol{H})$ is defined as the smallest integer k such that there exist k linearly dependent columns in $\boldsymbol{H}$ [23]. We will adopt the assumption (6) throughout the paper. Similarly, in the problem (1) we will assume that

$$\operatorname{spark}(\boldsymbol{D}) > 2s. \tag{7}$$

B. Estimation Techniques

We now review some standard estimators for the sparse problems described above. These techniques are usually viewed as methods for obtaining an estimate $\widehat{\boldsymbol{\alpha}}$ of the vector $\boldsymbol{\alpha}_0$ in (4), and we will adopt this perspective in the current section. One way to estimate $\boldsymbol{x}_0$ in the more general problem (1) is to first estimate $\boldsymbol{\alpha}_0$ with the methods described below and then use the formula $\widehat{\boldsymbol{x}} = \boldsymbol{D}\widehat{\boldsymbol{\alpha}}$.

A widely-used estimation technique is the ML approach, which provides an estimate of $\boldsymbol{\alpha}_0$ by solving

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\alpha}\|_2^2 \quad \text{s.t. } \|\boldsymbol{\alpha}\|_0 \leq s. \tag{8}$$

Unfortunately, (8) is a nonconvex optimization problem and solving it is NP-hard [21], meaning that an efficient algorithm providing the ML estimator is unlikely to exist. In fact, to the best of our knowledge, the most efficient method for solving (8) for general $\boldsymbol{H}$ is to enumerate the $\binom{p}{s}$ possible s-element support sets of $\boldsymbol{\alpha}$ and choose the one for which $\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{\alpha}\|_2^2$ is minimal. This is clearly an impractical strategy for reasonable values of p and s. Consequently, several efficient alternatives have been proposed for estimating $\boldsymbol{\alpha}_0$. One of these is the ℓ_1-penalty version of BPDN [1], which is defined as a solution $\widehat{\boldsymbol{\alpha}}_{\mathrm{BP}}$ to the quadratic program

$$\min_{\boldsymbol{\alpha}} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\alpha}\|_2^2 + \gamma\|\boldsymbol{\alpha}\|_1 \tag{9}$$

with some regularization parameter γ. More recently, the DS was proposed [4]; this approach estimates $\boldsymbol{\alpha}_0$ as a solution $\widehat{\boldsymbol{\alpha}}_{\mathrm{DS}}$ to

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \quad \text{s.t. } \|\boldsymbol{H}^T(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\alpha})\|_\infty \leq \tau \tag{10}$$

where τ is again a user-selected parameter. A modification of the DS, known as the Gauss–Dantzig selector (GDS) [4], is to use $\widehat{\boldsymbol{\alpha}}_{\mathrm{DS}}$ only to estimate the support of $\boldsymbol{\alpha}_0$. In this approach, one solves (10) and determines the support set of $\widehat{\boldsymbol{\alpha}}_{\mathrm{DS}}$. The GDS estimate is then obtained as

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{GDS}} = \begin{cases} \boldsymbol{H}_{\widehat{\boldsymbol{\alpha}}_{\mathrm{DS}}}^{\dagger}\boldsymbol{y} & \text{on the support set of } \widehat{\boldsymbol{\alpha}}_{\mathrm{DS}} \\ \boldsymbol{0} & \text{elsewhere} \end{cases} \tag{11}$$

where $\boldsymbol{H}_{\widehat{\boldsymbol{\alpha}}_{\mathrm{DS}}}$ consists of the columns of $\boldsymbol{H}$ corresponding to the support of $\widehat{\boldsymbol{\alpha}}_{\mathrm{DS}}$.

Previous research on the performance of these estimators has primarily examined their worst-case MSE among all possible values of $\boldsymbol{\alpha}_0 \in \mathcal{T}$. Specifically, it has been shown [4] that, under suitable conditions on $\boldsymbol{H}$, s, and τ, the DS of (10) satisfies

$$\|\boldsymbol{\alpha}_0 - \widehat{\boldsymbol{\alpha}}_{\mathrm{DS}}\|_2^2 \leq Cs\sigma^2 \log p \quad \text{with high probability} \tag{12}$$

for some constant C. It follows that the MSE of the DS is also no greater than a constant times $s\sigma^2 \log p$ for all $\boldsymbol{\alpha}_0 \in \mathcal{T}$ [12]. An identical property was also demonstrated for BPDN (9) with an appropriate choice of γ [11]. Conversely, it is known that the worst-case error of *any* estimator is at least a constant times $s\sigma^2 \log p$ [12, §7.4]. Thus, both BPDN and the DS are optimal, up to a constant, in terms of worst-case error. Nevertheless, the MSE of these approaches for specific values of $\boldsymbol{\alpha}_0$, even for a vast majority of such values, might be much lower. Our goal differs from this line of work in that we characterize the *pointwise* performance of an estimator, i.e., the MSE for specific values of $\boldsymbol{\alpha}_0$.

Another baseline with which practical techniques are often compared is the oracle estimator, given by

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{oracle}} = \begin{cases} \boldsymbol{H}_{\boldsymbol{\alpha}_0}^{\dagger}\boldsymbol{b} & \text{on the set } \operatorname{supp}(\boldsymbol{\alpha}_0) \\ \boldsymbol{0} & \text{elsewhere} \end{cases} \tag{13}$$

where $\boldsymbol{H}_{\boldsymbol{\alpha}_0}$ is the submatrix constructed from the columns of $\boldsymbol{H}$ corresponding to the nonzero entries of $\boldsymbol{\alpha}_0$. In other words, $\widehat{\boldsymbol{\alpha}}_{\mathrm{oracle}}$ is the least-squares (LS) solution among vectors whose support coincides with $\operatorname{supp}(\boldsymbol{\alpha}_0)$, which is assumed to have been provided by an "oracle." Of course, in practice the support of $\boldsymbol{\alpha}_0$ is unknown, so that $\widehat{\boldsymbol{\alpha}}_{\mathrm{oracle}}$ cannot actually be implemented. Nevertheless, one often compares the performance of true estimators with $\widehat{\boldsymbol{\alpha}}_{\mathrm{oracle}}$, whose MSE is given by [4]

$$\sigma^2 \operatorname{Tr}((\boldsymbol{H}_{\boldsymbol{\alpha}_0}^T\boldsymbol{H}_{\boldsymbol{\alpha}_0})^{-1}). \tag{14}$$

Is (14) a bound on estimation MSE? While $\widehat{\boldsymbol{\alpha}}_{\mathrm{oracle}}$ is a reasonable technique to adopt if $\operatorname{supp}(\boldsymbol{\alpha}_0)$ is known, this does not imply that (14) is a lower bound on the performance of practical estimators. Indeed, as will be demonstrated in Section V, when the SNR is low, both BPDN and the DS outperform $\widehat{\boldsymbol{\alpha}}_{\mathrm{oracle}}$, thanks to the use of shrinkage in these estimators. Furthermore, if $\operatorname{supp}(\boldsymbol{\alpha}_0)$ is known, then there exist biased techniques which are better than $\widehat{\boldsymbol{\alpha}}_{\mathrm{oracle}}$ for *all* values of $\boldsymbol{\alpha}_0$ [24]. Thus, $\widehat{\boldsymbol{\alpha}}_{\mathrm{oracle}}$ is neither achievable in practice, nor optimal in terms of MSE. As we will see, one can indeed interpret (14) as a lower bound on the achievable MSE, but such a result requires a certain restriction of the class of estimators under consideration.

III. The Constrained Cramér–Rao Bound

A common technique for determining the achievable performance in a given estimation problem is to calculate the CRB, which is a lower bound on the MSE of estimators having a given bias [13]. In this paper, we are interested in calculating the CRB when it is known that the parameter $\boldsymbol{x}$ satisfies sparsity constraints such as those of the sets $\mathcal{S}$ of (2) and $\mathcal{T}$ of (5).

The CRB for constrained parameter sets has been studied extensively in the past [15]–[18]. However, in prior work derivation of the CRB assumed that the constraint set is given by

$$\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0},\ \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}\} \tag{15}$$

where $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x})$ are continuously differentiable functions. We will refer to such $\mathcal{X}$ as continuously differentiable sets. As shown in prior work [15], the resulting bound depends on the derivatives of the function $\boldsymbol{f}$. Yet in some cases, including the sparse estimation scenarios discussed in Section II, the constraint set cannot be written in the form (15), and the aforementioned results are therefore inapplicable. Our goal in the current section is to close this gap by extending the constrained CRB to constraint sets $\mathcal{X}$ encompassing the sparse estimation scenario.

We begin this section with a general discussion of the CRB and the class of estimators to which it applies. This will lead us to interpret the constrained CRB as a bound on estimators having an incompletely specified bias gradient. This interpretation will facilitate the application of the existing constrained CRB to the present context.

A. Bias Requirements in the Constrained CRB

In previous settings for which the constrained CRB was derived, it was noted that the resulting bound is typically lower than the unconstrained version [15, Remark 4]. At first glance, one would attribute the reduction in the value of the CRB to the fact that the constraints add information about the unknown parameter, which can then improve estimation performance. On the other hand, the CRB separately characterizes the achievable performance for each value of the unknown parameter $\boldsymbol{x}_0$. Thus, the CRB at $\boldsymbol{x}_0$ applies even to estimators designed specifically to perform well at $\boldsymbol{x}_0$. Such estimators surely cannot achieve further gain in performance if it is known that $\boldsymbol{x}_0 \in \mathcal{X}$. Why, then, is the constrained CRB lower than the unconstrained bound? The answer to this apparent paradox involves a careful definition of the class of estimators to which the bound applies.

To obtain a meaningful bound, one must exclude some estimators from consideration. Unless this is done, the bound will be tarnished by estimators of the type $\widehat{\boldsymbol{x}} = \boldsymbol{x}_{\mathrm{u}}$, for some constant $\boldsymbol{x}_{\mathrm{u}}$, which achieve an MSE of 0 at the specific point $\boldsymbol{x} = \boldsymbol{x}_{\mathrm{u}}$. It is standard practice to circumvent this difficulty by restricting attention to estimators having a particular bias $\boldsymbol{b}(\boldsymbol{x}) \triangleq E\{\widehat{\boldsymbol{x}}\} - \boldsymbol{x}$. In particular, it is common to examine unbiased estimators, for which $\boldsymbol{b}(\boldsymbol{x}) = \boldsymbol{0}$.

However, in some settings, it is impossible to construct estimators which are unbiased for all $\boldsymbol{x} \in \mathbb{R}^n$. For example, suppose we are to estimate the coefficients $\boldsymbol{\alpha}_0$ of an overcomplete dictionary based on the measurements given by (4). Since the dictionary is overcomplete, its nullspace is nontrivial; furthermore, each coefficient vector in the nullspace yields an identical distribution of the measurements, so that an estimator can be unbiased for one of these vectors at most.

The question is whether it is possible to construct estimators which are unbiased for some, but not all, values of $\boldsymbol{x}$. One possible approach is to seek estimators which are unbiased for all $\boldsymbol{x} \in \mathcal{X}$. However, as we will see later in this section, even this requirement can be too strict: in some cases it is impossible to construct estimators which are unbiased for all $\boldsymbol{x} \in \mathcal{X}$. More generally, the CRB is a *local* bound, meaning that it determines the achievable performance at a particular value of $\boldsymbol{x}$ based on the statistics at $\boldsymbol{x}$ and at nearby values. Thus, it is irrelevant to introduce requirements on estimation performance for parameters which are distant from the value $\boldsymbol{x}$ of interest.

Since we seek a locally unbiased estimator, one possibility is to require unbiasedness at a single point, say $\boldsymbol{x}_{\mathrm{u}}$. As it turns out, it is always possible to construct such a technique: this is again $\widehat{\boldsymbol{x}} = \boldsymbol{x}_{\mathrm{u}}$, which is unbiased at $\boldsymbol{x}_{\mathrm{u}}$ but nowhere else. To avoid this loophole, one can require an estimator to be unbiased in the neighborhood

$$\mathcal{B}_\varepsilon(\boldsymbol{x}_0) = \{\boldsymbol{x} \in \mathbb{R}^m : \|\boldsymbol{x} - \boldsymbol{x}_0\|_2 < \varepsilon\} \tag{16}$$

of $\boldsymbol{x}_0$, for some small ε. It follows that both the bias $\boldsymbol{b}(\boldsymbol{x})$ and the bias gradient

$$\boldsymbol{B}(\boldsymbol{x}) \triangleq \frac{\partial \boldsymbol{b}}{\partial \boldsymbol{x}} \tag{17}$$

vanish at $\boldsymbol{x} = \boldsymbol{x}_0$. This formulation is the basis of the unconstrained unbiased CRB, a lower bound on the covariance at $\boldsymbol{x}_0$ which applies to all estimators whose bias gradient is zero at $\boldsymbol{x}_0$.

It turns out that even this requirement is too stringent in constrained settings. As we will see in Section IV-A, estimators of the coefficients of an overcomplete dictionary must have a nonzero bias gradient matrix. The reason is related to the fact that unbiasedness is required over the set $\mathcal{B}_\varepsilon(\boldsymbol{x}_0)$, which, in the overcomplete setting, has a higher dimension than the number of measurements.

However, it can be argued that one is not truly interested in the bias at all points in $\mathcal{B}_\varepsilon(\boldsymbol{x}_0)$, since many of these points violate the constraint set $\mathcal{X}$. A reasonable compromise is to require unbiasedness over $\mathcal{B}_\varepsilon(\boldsymbol{x}_0) \cap \mathcal{X}$, i.e., over the neighborhood of $\boldsymbol{x}_0$ restricted to the constraint set $\mathcal{X}$. This leads to a weaker requirement on the bias gradient $\boldsymbol{B}$ at $\boldsymbol{x}_0$. Specifically, the derivatives of the bias need only be specified in directions which do not violate the constraints. The exact formulation of this requirement depends on the nature of the set $\mathcal{X}$. In the following subsections, we will investigate various constraint sets and derive the corresponding requirements on the bias function.

It is worth emphasizing that the dependence of the CRB on the constraints is manifested through the class of estimators being considered, or more specifically, through the allowed estimators' bias gradient matrices. By contrast, the unconstrained CRB applies to estimators having a fully specified bias

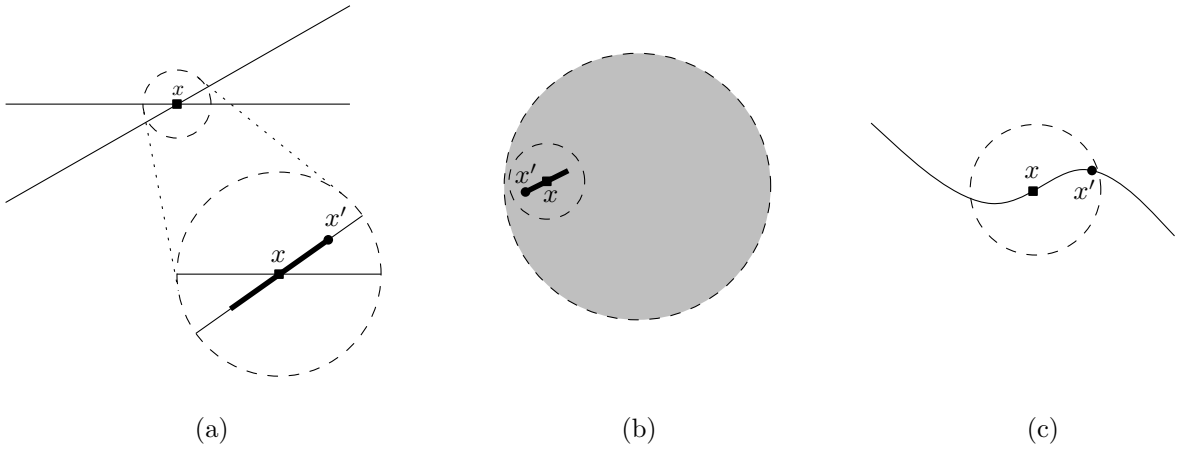

Fig. 1. In a locally balanced set such as a union of subspaces (a) and an open ball (b), each point is locally defined by a set of feasible directions along which an infinitesimal movement does not violate the constraints. The curve (c) is not characterized in this way and thus is not locally balanced.

gradient matrix. Consequently, the constrained bound applies to a wider class of estimators, and is thus usually lower than the unconstrained version of the CRB. In other words, estimators which are unbiased in the constrained setting, and thus applicable to the unbiased constrained CRB, are likely to be biased in the unconstrained context. Since a wider class of estimators is considered by the constrained CRB, the resulting bound is lower, thus explaining the puzzling phenomenon described in the beginning of this subsection.

B. Locally Balanced Constraints

We now consider a class of constraint sets, called locally balanced sets, which encompass the sparsity constraints of Section II. Roughly speaking, a locally balanced set is one which is locally defined at each point by the directions along which one can move without leaving the set. Formally, a metric space $\mathcal{X}$ is said to be locally balanced if, for all $\boldsymbol{x} \in \mathcal{X}$, there exists an open set $\mathcal{C} \subset \mathcal{X}$ such that $\boldsymbol{x} \in \mathcal{C}$ and such that, for all $\boldsymbol{x}' \in \mathcal{C}$ and for all $|\lambda| \leq 1$, we have

$$\boldsymbol{x} + \lambda(\boldsymbol{x}' - \boldsymbol{x}) \in \mathcal{C}. \tag{18}$$

As we will see, locally balanced sets are useful in the context of the constrained CRB, as they allow us to identify the feasible directions along which the bias gradient must be specified.

An example of a locally balanced set is given in Fig. 1(a), which represents a union of two subspaces. In Fig. 1(a), for any point $\boldsymbol{x} \in \mathcal{X}$, and for any point $\boldsymbol{x}' \in \mathcal{X}$ sufficiently close to $\boldsymbol{x}$, the entire line segment between $\boldsymbol{x}$ and $\boldsymbol{x}'$, as well as the line segment in the opposite direction, are also in $\mathcal{X}$. This illustrates the fact that any union of subspaces is locally balanced, and, in particular, so are the sparse estimation settings of Section II [25]–[27]. As another example, consider any open set, such as the open ball in Fig. 1(b). For such a set, any point $\boldsymbol{x}$ has a sufficiently small neighborhood $\mathcal{C}$ such that, for any $\boldsymbol{x}' \in \mathcal{C}$, the line segment connecting $\boldsymbol{x}$ to $\boldsymbol{x}'$ is contained in $\mathcal{X}$. On the other hand, the curve in Fig. 1(c) is not locally balanced, since the line connecting $\boldsymbol{x}$ to any other point on the set does not lie within the set.[1]

Observe that the neighborhood of a point $\boldsymbol{x}$ in a locally balanced set $\mathcal{X}$ is entirely determined by the set of feasible directions $\boldsymbol{v}$ along which infinitesimal changes of $\boldsymbol{x}$ do not violate the constraints. These are the directions $\boldsymbol{v} = \boldsymbol{x}' - \boldsymbol{x}$ for all points $\boldsymbol{x}' \neq \boldsymbol{x}$ in the set $\mathcal{C}$ of (18). Recall that we seek a lower bound on the performance of estimators whose bias gradient is defined over the neighborhood of $\boldsymbol{x}_0$ restricted to the constraint set $\mathcal{X}$. Suppose for concreteness that we are interested in unbiased estimators. For a locally balanced constraint set $\mathcal{X}$, this implies that

$$\boldsymbol{B}\boldsymbol{v} = \boldsymbol{0} \tag{19}$$

for any feasible direction $\boldsymbol{v}$. In other words, all feasible directions must be in the nullspace of $\boldsymbol{B}$. This is a weaker condition than requiring the bias gradient to equal zero, and is thus more useful for constrained estimation problems. If an estimator $\widehat{\boldsymbol{x}}$ satisfies (19) for all feasible directions $\boldsymbol{v}$ at a certain point $\boldsymbol{x}_0$, we say that $\widehat{\boldsymbol{x}}$ is $\mathcal{X}$-unbiased at $\boldsymbol{x}_0$. This terminology emphasizes the fact that $\mathcal{X}$-unbiasedness depends both on the point $\boldsymbol{x}_0$ and on the constraint set $\mathcal{X}$.

Consider the subspace $\mathcal{F}$ spanned by the feasible directions at a certain point $\boldsymbol{x} \in \mathcal{X}$. We refer to $\mathcal{F}$ as the feasible subspace at $\boldsymbol{x}$. Note that $\mathcal{F}$ may include infeasible directions, if these are linear combinations of feasible directions. Nevertheless, because of the linearity of (19), any vector $\boldsymbol{u} \in \mathcal{F}$ satisfies $\boldsymbol{B}\boldsymbol{u} = \boldsymbol{0}$, even if $\boldsymbol{u}$ is infeasible. Thus, $\mathcal{X}$-unbiasedness is actually a property of the feasible subspace $\mathcal{F}$, rather than the set of feasible directions.

Since $\mathcal{X}$ is a subset of a finite-dimensional Euclidean space, $\mathcal{F}$ is also finite-dimensional, although different points in $\mathcal{X}$ may yield subspaces having differing dimensions. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l$ denote an orthonormal basis for $\mathcal{F}$, and define the matrix

$$\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l]. \tag{20}$$

[1]We note in passing that since the curve in Fig. 1(c) is continuously differentiable, it can be locally approximated by a locally balanced set. Our derivation of the CRB can be extended to such approximately locally balanced sets in a manner similar to that of [15], but such an extension is not necessary for the purposes of this paper.

Note that $\boldsymbol{u}_i$ and $\boldsymbol{U}$ are functions of $\boldsymbol{x}$. For a given function $\boldsymbol{x}$, different orthonormal bases can be chosen, but the choice of a basis is arbitrary and will not affect our results.

As we have seen, $\mathcal{X}$-unbiasedness at $\boldsymbol{x}_0$ can alternatively be written as $\boldsymbol{B}\boldsymbol{u} = \boldsymbol{0}$ for all $\boldsymbol{u} \in \mathcal{F}$, or, equivalently

$$\boldsymbol{B}\boldsymbol{U} = \boldsymbol{0}. \tag{21}$$

The constrained CRB can now be derived as a lower bound on all $\mathcal{X}$-unbiased estimators, which is a weaker requirement than "ordinary" unbiasedness.

Just as $\mathcal{X}$-unbiasedness was defined by requiring the bias gradient matrix to vanish when multiplied by any feasible direction vector, we can define $\mathcal{X}$-biased estimators by requiring a specific value (not necessarily zero) for the bias gradient matrix when multiplied by a feasible direction vector. In an analogy to (21), this implies that one must define a value for the matrix $\boldsymbol{B}\boldsymbol{U}$. Our goal is thus to construct a lower bound on the covariance at a given $\boldsymbol{x}$ achievable by any estimator whose bias gradient $\boldsymbol{B}$ at $\boldsymbol{x}$ satisfies $\boldsymbol{B}\boldsymbol{U} = \boldsymbol{P}$, for a given matrix $\boldsymbol{P}$. This is referred to as specifying the $\mathcal{X}$-bias of the estimator at $\boldsymbol{x}$.

C. The CRB for Locally Balanced Constraints

It is helpful at this point to compare our derivation with prior work on the constrained CRB, which considered continuously differentiable constraint sets of the form (15). It has been previously shown [15] that inequality constraints of the type $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$ have no effect on the CRB. Consequently, we will consider constraints of the form

$$\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}\}. \tag{22}$$

Define the $k \times n$ matrix $\boldsymbol{F}(\boldsymbol{x}) = \partial \boldsymbol{f}/\partial \boldsymbol{x}$. For simplicity of notation, we will omit the dependence of $\boldsymbol{F}$ on $\boldsymbol{x}$. Assuming that the constraints are non-redundant, $\boldsymbol{F}$ is a full-rank matrix, and thus one can define an $n \times (n-k)$ matrix $\boldsymbol{W}$ (also dependent on $\boldsymbol{x}$) such that

$$\boldsymbol{F}\boldsymbol{W} = \boldsymbol{0}, \quad \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}. \tag{23}$$

The matrix $\boldsymbol{W}$ is closely related to the matrix $\boldsymbol{U}$ spanning the feasible direction subspace of locally balanced sets. Indeed, the column space $\mathcal{R}(\boldsymbol{W})$ of $\boldsymbol{W}$ is the tangent space of $\mathcal{X}$, i.e., the subspace of $\mathbb{R}^n$ containing all vectors which are tangent to $\mathcal{X}$ at the point $\boldsymbol{x}$. Thus, the vectors in $\mathcal{R}(\boldsymbol{W})$ are precisely those directions along which infinitesimal motion from $\boldsymbol{x}$ does not violate the constraints, up to a first-order approximation. It follows that if a particular set $\mathcal{X}$ is both locally balanced and continuously differentiable, its matrices $\boldsymbol{U}$ and $\boldsymbol{W}$ coincide. Note, however, that there exist sets which are locally balanced but not continuously differentiable (and vice versa).

With the above formulation, the CRB for continuously differentiable constraints can be stated as a function of the the matrix $\boldsymbol{W}$ and the bias gradient $\boldsymbol{B}$ [18]. In fact, the resulting bound depends on $\boldsymbol{B}$ only through $\boldsymbol{B}\boldsymbol{W}$. This is to be expected in light of the discussion of Section III-A: The bias should be specified only for those directions which do not violate the constraint set. Furthermore, the proof of the CRB in [18, Theorem 1] depends not on the formulation (22) of the constraint set, but merely on the class of bias functions under consideration. Consequently, one can state the bound without any reference to the underlying constraint set. To do so, let $\boldsymbol{y}$ be a measurement vector with pdf $p(\boldsymbol{y};\boldsymbol{x})$, which is assumed to be differentiable with respect to $\boldsymbol{x}$. The Fisher information matrix (FIM) $\boldsymbol{J}(\boldsymbol{x})$ is defined as

$$\boldsymbol{J}(\boldsymbol{x}) = E\Big\{\boldsymbol{\Delta}\boldsymbol{\Delta}^T\Big\} \tag{24}$$

where

$$\boldsymbol{\Delta} = \frac{\partial \log p(\boldsymbol{y};\boldsymbol{x})}{\partial \boldsymbol{x}}. \tag{25}$$

We assume that the FIM is well-defined and finite. We further assume that integration with respect to $\boldsymbol{y}$ and differentiation with respect to $\boldsymbol{x}$ can be interchanged, a standard requirement for the CRB. We then have the following result.

Theorem 1: Let $\widehat{\boldsymbol{x}}$ be an estimator and let $\boldsymbol{B} = \partial \boldsymbol{b}/\partial \boldsymbol{x}$ denote the bias gradient matrix of $\widehat{\boldsymbol{x}}$ at a given point $\boldsymbol{x}_0$. Let $\boldsymbol{U}$ be an orthonormal matrix, and suppose that $\boldsymbol{B}\boldsymbol{U}$ is known, but that $\boldsymbol{B}$ is otherwise arbitrary. If

$$\mathcal{R}(\boldsymbol{U}(\boldsymbol{U}+\boldsymbol{B}\boldsymbol{U})^T)) \subseteq \mathcal{R}(\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{J}\boldsymbol{U}\boldsymbol{U}^T) \tag{26}$$

then the covariance of $\widehat{\boldsymbol{x}}$ at $\boldsymbol{x}_0$ satisfies

$$\mathrm{Cov}(\widehat{\boldsymbol{x}}) \succeq (\boldsymbol{U}+\boldsymbol{B}\boldsymbol{U})\left(\boldsymbol{U}^T\boldsymbol{J}\boldsymbol{U}\right)^{\dagger}(\boldsymbol{U}+\boldsymbol{B}\boldsymbol{U})^T. \tag{27}$$

Equality is achieved in (27) if and only if

$$\widehat{\boldsymbol{x}} = \boldsymbol{x}_0 + \boldsymbol{b}(\boldsymbol{x}_0) + (\boldsymbol{U}+\boldsymbol{B}\boldsymbol{U})\left(\boldsymbol{U}^T\boldsymbol{J}\boldsymbol{U}\right)^{\dagger}\boldsymbol{U}^T\boldsymbol{\Delta} \tag{28}$$

in the mean square sense, where $\boldsymbol{\Delta}$ is defined by (25). Conversely, if (26) does not hold, then there exists no finite-variance estimator with the required bias gradient.

As required, no mention of constrained estimation is made in Theorem 1; instead, partial information about the bias gradient is assumed. Apart from this restatement, the theorem is identical to [18, Theorem 1], and its proof is unchanged. However, the above formulation is more general in that it can be applied to any constrained setting, once the constraints have been translated to bias gradient requirements. In particular, Theorem 1 provides a CRB for locally balanced sets if the matrix $\boldsymbol{U}$ is chosen as a basis for the feasible direction subspace of Section III-B.

IV. Bounds on Sparse Estimation

In this section, we apply the CRB of Theorem 1 to several sparse estimation scenarios. We begin with an analysis of the problem of estimating a sparse parameter vector.

A. Estimating a Sparse Vector

Suppose we would like to estimate a parameter vector $\boldsymbol{\alpha}_0$, known to belong to the set $\mathcal{T}$ of (5), from measurements $\boldsymbol{y}$ given by (4). To determine the CRB in this setting, we begin by identifying the feasible subspaces $\mathcal{F}$ corresponding to each of the elements in $\mathcal{T}$. To this end, consider first vectors $\boldsymbol{\alpha} \in \mathcal{T}$ for which $\|\boldsymbol{\alpha}\|_0 = s$, i.e., vectors having maximal support. Denote by $\{i_1, \ldots, i_s\}$ the support set of $\boldsymbol{\alpha}$. Then, for all δ, we have

$$\|\boldsymbol{\alpha} + \delta \boldsymbol{e}_{i_k}\|_0 = \|\boldsymbol{\alpha}\|_0 = s, \quad k = 1, \ldots, s \tag{29}$$

where $\boldsymbol{e}_j$ is the jth column of the identity matrix. Thus $\boldsymbol{\alpha} + \delta\boldsymbol{e}_{i_k} \in \mathcal{T}$, and consequently, the vectors $\{\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_s}\}$ are all feasible directions, as is any linear combination of these vectors. On the other hand, for any $j \notin \mathrm{supp}(\boldsymbol{\alpha})$ and for any nonzero δ, we have $\|\boldsymbol{\alpha} + \delta\boldsymbol{e}_j\|_0 = s+1$, and thus $\boldsymbol{e}_j$ is not a feasible direction; neither is any other vector which is not in $\mathrm{span}\{\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_s}\}$. It follows that the feasible subspace $\mathcal{F}$ for points having maximal support is given by $\mathrm{span}\{\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_s}\}$, and a possible choice for the matrix $\boldsymbol{U}$ of (20) is

$$\boldsymbol{U} = [\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_s}] \quad \text{for } \|\boldsymbol{\alpha}\|_0 = s. \tag{30}$$

The situation is different for points $\boldsymbol{\alpha}$ having $\|\boldsymbol{\alpha}\|_0 < s$. In this case, vectors $\boldsymbol{e}_i$ corresponding to *any* direction i are feasible directions, since

$$\|\boldsymbol{\alpha} + \delta\boldsymbol{e}_i\|_0 \leq \|\boldsymbol{\alpha}\|_0 + 1 \leq s. \tag{31}$$

Because the feasible subspace is defined as the span of all feasible directions, we have

$$\mathcal{F} \supseteq \mathrm{span}\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p\} = \mathbb{R}^p. \tag{32}$$

It follows that $\mathcal{F} = \mathbb{R}^p$ and thus a convenient choice for the matrix $\boldsymbol{U}$ is

$$\boldsymbol{U} = \boldsymbol{I} \quad \text{for } \|\boldsymbol{\alpha}\|_0 < s. \tag{33}$$

Consequently, whenever $\|\boldsymbol{\alpha}\|_0 < s$, a specification of the $\mathcal{T}$-bias amounts to completely specifying the usual estimation bias $\boldsymbol{b}(\boldsymbol{x})$.

To invoke Theorem 1, we must also determine the FIM $\boldsymbol{J}(\boldsymbol{\alpha})$. Under our assumption of white Gaussian noise, $\boldsymbol{J}(\boldsymbol{\alpha})$ is given by [13, p. 85]

$$\boldsymbol{J}(\boldsymbol{\alpha}) = \frac{1}{\sigma^2}\boldsymbol{H}^T\boldsymbol{H}. \tag{34}$$

Using (30), (33), and (34), it is readily shown that

$$\boldsymbol{U}^T\boldsymbol{J}\boldsymbol{U} = \begin{cases} \frac{1}{\sigma^2}\boldsymbol{H}_{\boldsymbol{\alpha}}^T\boldsymbol{H}_{\boldsymbol{\alpha}} & \text{when } \|\boldsymbol{\alpha}\|_0 = s \\ \frac{1}{\sigma^2}\boldsymbol{H}^T\boldsymbol{H} & \text{when } \|\boldsymbol{\alpha}\|_0 < s \end{cases} \tag{35}$$

where $\boldsymbol{H}_{\boldsymbol{\alpha}}$ is the $p \times s$ matrix consisting of the columns of $\boldsymbol{H}$ indexed by $\mathrm{supp}(\boldsymbol{\alpha})$.

We now wish to determine under what conditions (26) holds. Consider first points $\boldsymbol{\alpha}_0$ for which $\|\boldsymbol{\alpha}_0\|_0 = s$. Since, by (6), we have $\mathrm{spark}(\boldsymbol{H}) > s$, it follows that in this case $\boldsymbol{U}^T\boldsymbol{J}\boldsymbol{U}$ is invertible. Therefore

$$\mathcal{R}(\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{J}\boldsymbol{U}\boldsymbol{U}^T) = \mathcal{R}(\boldsymbol{U}\boldsymbol{U}^T). \tag{36}$$

Since

$$\mathcal{R}(\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{I} + \boldsymbol{B}^T)) \subseteq \mathcal{R}(\boldsymbol{U}\boldsymbol{U}^T) \tag{37}$$

we have that condition (26) holds when $\|\boldsymbol{\alpha}_0\|_0 = s$.

The condition (26) is no longer guaranteed when $\|\boldsymbol{\alpha}_0\|_0 < s$. In this case, $\boldsymbol{U} = \boldsymbol{I}$, so that (26) is equivalent to

$$\mathcal{R}(\boldsymbol{I} + \boldsymbol{B}^T) \subseteq \mathcal{R}(\boldsymbol{H}^T\boldsymbol{H}). \tag{38}$$

Using the fact that $\mathcal{R}(\boldsymbol{H}^T\boldsymbol{H}) = \mathcal{R}(\boldsymbol{H}^T)$ and that, for any matrix $\boldsymbol{Q}$, $\mathcal{R}(\boldsymbol{Q}^T) = \mathcal{N}(\boldsymbol{Q})^\perp$, we find that (38) is equivalent to

$$\mathcal{N}(\boldsymbol{H}) \subseteq \mathcal{N}(\boldsymbol{I} + \boldsymbol{B}). \tag{39}$$

Combining these conclusions with Theorem 1 yields the following CRB for the problem of estimating a sparse vector.

Theorem 2: Consider the estimation problem (4) with $\boldsymbol{\alpha}_0$ given by (5), and assume that (6) holds. For a finite-variance estimator $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}_0$ to exist, its bias gradient matrix $\boldsymbol{B}$ must satisfy (39) whenever $\|\boldsymbol{\alpha}_0\|_0 < s$. Furthermore, the covariance of any estimator whose $\mathcal{T}$-bias gradient matrix is $\boldsymbol{B}\boldsymbol{U}$ satisfies

$$\begin{aligned} \mathrm{Cov}(\widehat{\boldsymbol{\alpha}}) &\succeq \sigma^2(\boldsymbol{I} + \boldsymbol{B})(\boldsymbol{H}^T\boldsymbol{H})^\dagger(\boldsymbol{I} + \boldsymbol{B}^T) \\ &\qquad \text{when } \|\boldsymbol{\alpha}_0\|_0 < s, \\ \mathrm{Cov}(\widehat{\boldsymbol{\alpha}}) &\succeq \sigma^2(\boldsymbol{U} + \boldsymbol{B}\boldsymbol{U})(\boldsymbol{H}_{\boldsymbol{\alpha}_0}^T\boldsymbol{H}_{\boldsymbol{\alpha}_0})^{-1}(\boldsymbol{U} + \boldsymbol{B}\boldsymbol{U})^T \\ &\qquad \text{when } \|\boldsymbol{\alpha}_0\|_0 = s. \end{aligned} \tag{40}$$

Here, $\boldsymbol{H}_{\boldsymbol{\alpha}_0}$ is the matrix containing the columns of $\boldsymbol{H}$ corresponding to $\mathrm{supp}(\boldsymbol{\alpha}_0)$.

Let us examine Theorem 2 separately in the underdetermined and well-determined cases. In the well-determined case, in which $\boldsymbol{H}$ has full row rank, the nullspace of $\boldsymbol{H}$ is trivial, so that (39) always holds. It follows that the CRB is always finite, in the sense that we cannot rule out the existence of an estimator having any given bias function. Some insight can be obtained in this case by examining the $\mathcal{T}$-unbiased case. Noting also that $\boldsymbol{H}^T\boldsymbol{H}$ is invertible in the well-determined case, the bound for $\mathcal{T}$-unbiased estimators is given by

$$\begin{aligned} \mathrm{Cov}(\widehat{\boldsymbol{\alpha}}) &\succeq \sigma^2(\boldsymbol{H}^T\boldsymbol{H})^{-1} && \text{when } \|\boldsymbol{\alpha}_0\|_0 < s, \\ \mathrm{Cov}(\widehat{\boldsymbol{\alpha}}) &\succeq \sigma^2\boldsymbol{U}(\boldsymbol{H}_{\boldsymbol{\alpha}_0}^T\boldsymbol{H}_{\boldsymbol{\alpha}_0})^{-1}\boldsymbol{U}^T && \text{when } \|\boldsymbol{\alpha}_0\|_0 = s. \end{aligned} \tag{41}$$

From this formulation, the behavior of the CRB can be described as follows. When $\boldsymbol{\alpha}_0$ has non-maximal support ($\|\boldsymbol{\alpha}_0\|_0 < s$), the CRB is identical to the bound which would have been obtained had there been no constraints in the problem. This is because $\boldsymbol{U} = \boldsymbol{I}$ in this case, so that $\mathcal{T}$-unbiasedness and ordinary unbiasedness are equivalent. As we have seen in Section III-A, the CRB is a function of the class of estimators under consideration, so the unconstrained and constrained bounds are equivalent in this situation. The bound $\sigma^2(\boldsymbol{H}^T\boldsymbol{H})^{-1}$ is achieved by the unconstrained LS estimator

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{y} \tag{42}$$

which is the minimum variance unbiased estimator in the unconstrained case. Thus, we learn from Theorem 2 that for values of $\boldsymbol{\alpha}_0$ having non-maximal support, no $\mathcal{T}$-unbiased technique can outperform the standard LS estimator, which does not assume any knowledge about the constraint set $\mathcal{T}$.

On the other hand, consider the case in which $\boldsymbol{\alpha}_0$ has maximal support, i.e., $\|\boldsymbol{\alpha}_0\|_0 = s$. Suppose first that $\mathrm{supp}(\boldsymbol{\alpha}_0)$ is known, so that one must estimate only the nonzero values of $\boldsymbol{\alpha}_0$. In this case, a reasonable approach is to use the oracle estimator (13), whose covariance matrix is given by $\sigma^2\boldsymbol{U}(\boldsymbol{H}_{\boldsymbol{\alpha}_0}^T\boldsymbol{H}_{\boldsymbol{\alpha}_0})^{-1}\boldsymbol{U}^T$ [4]. Thus, when $\boldsymbol{\alpha}_0$ has maximal support, Theorem 2 states that $\mathcal{T}$-unbiased estimators can perform, at best, as well as the oracle estimator, which is equivalent to the LS approach when the support of $\boldsymbol{\alpha}_0$ is known.

The situation is similar, but somewhat more involved, in the underdetermined case. Here, the condition (39) for the

existence of an estimator having a given bias gradient matrix no longer automatically holds. To interpret this condition, it is helpful to introduce the mean gradient matrix $\boldsymbol{M}(\boldsymbol{\alpha})$, defined as

$$\boldsymbol{M}(\boldsymbol{\alpha}) = \frac{\partial E\{\widehat{\boldsymbol{\alpha}}\}}{\partial \boldsymbol{\alpha}} = \boldsymbol{I} + \boldsymbol{B}. \tag{43}$$

The matrix $\boldsymbol{M}(\boldsymbol{\alpha})$ is a measure of the sensitivity of an estimator to changes in the parameter vector. For example, a $\mathcal{T}$-unbiased estimator is sensitive to any *feasible* change in $\boldsymbol{\alpha}$. Thus, $\mathcal{N}(\boldsymbol{M})$ denotes the subspace of directions to which $\widehat{\boldsymbol{\alpha}}$ is insensitive. Likewise, $\mathcal{N}(\boldsymbol{H})$ is the subspace of directions for which a change in $\boldsymbol{\alpha}$ does not modify $\boldsymbol{H}\boldsymbol{\alpha}$. The condition (39) therefore states that for an estimator to exist, it must be insensitive to changes in $\boldsymbol{\alpha}$ which are unobservable through $\boldsymbol{H}\boldsymbol{\alpha}$, at least when $\|\boldsymbol{\alpha}\|_0 < s$. No such requirement is imposed in the case $\|\boldsymbol{\alpha}\|_0 = s$, since in this case there are far fewer feasible directions.

The lower bound (40) is similarly a consequence of the wide range of feasible directions obtained when $\|\boldsymbol{\alpha}\|_0 < s$, as opposed to the tight constraints when $\|\boldsymbol{\alpha}\|_0 = s$. Specifically, when $\|\boldsymbol{\alpha}\|_0 < s$, a change to any component of $\boldsymbol{\alpha}$ is feasible and hence the lower bound equals that of an unconstrained estimation problem, with the FIM given by $\sigma^{-2}\boldsymbol{H}^T\boldsymbol{H}$. On the other hand, when $\|\boldsymbol{\alpha}\|_0 = s$, the bound is effectively that of an estimator with knowledge of the particular subspace to which $\boldsymbol{\alpha}$ belongs; for this subspace the FIM is the submatrix $\boldsymbol{U}^T\boldsymbol{J}\boldsymbol{U}$ given in (35). This phenomenon is discussed further in Section VI.

Another difference between the well-determined and underdetermined cases is that when $\boldsymbol{H}$ is underdetermined, an estimator cannot be $\mathcal{T}$-unbiased for all $\boldsymbol{\alpha}$. To see this, recall from (21) that $\mathcal{T}$-unbiased estimators are defined by the fact that $\boldsymbol{B}\boldsymbol{U} = \boldsymbol{0}$. When $\|\boldsymbol{\alpha}\|_0 < s$, we have $\boldsymbol{U} = \boldsymbol{I}$ and thus $\mathcal{T}$-unbiasedness implies $\boldsymbol{B} = \boldsymbol{0}$, so that $\mathcal{N}(\boldsymbol{I}+\boldsymbol{B}) = \{\boldsymbol{0}\}$. But since $\boldsymbol{H}$ is underdetermined, $\mathcal{N}(\boldsymbol{H})$ is nontrivial. Consequently, (39) cannot hold for $\mathcal{T}$-unbiased estimators when $\|\boldsymbol{\alpha}\|_0 < s$.

The lack of $\mathcal{T}$-unbiased estimators when $\|\boldsymbol{\alpha}_0\|_0 < s$ is a direct consequence of the fact that the feasible direction set at such $\boldsymbol{\alpha}_0$ contains all of the directions $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$. The conclusion from Theorem 2 is then that no estimator can be expected to be unbiased in such a high-dimensional neighborhood, just as unbiased estimation is impossible in the p-dimensional neighborhood $\mathcal{B}_\varepsilon(\boldsymbol{\alpha}_0)$, as explained in Section III-A. However, it is still possible to obtain a finite CRB in this setting by further restricting the constraint set: if it is known that $\|\boldsymbol{\alpha}_0\|_0 = \tilde{s} < s$, then one can redefine $\mathcal{T}$ in (5) by replacing s with $\tilde{s}$. This will enlarge the class of estimators considered $\mathcal{T}$-unbiased, and Theorem 2 would then provide a finite lower bound on those estimators. Such estimators will not, however, be unbiased in the sense implied by the original constraint set.

While an estimator cannot be unbiased for *all* $\boldsymbol{\alpha} \in \mathcal{T}$, unbiasedness is possible at points $\boldsymbol{\alpha}$ for which $\|\boldsymbol{\alpha}\|_0 = s$. In this case, Theorem 2 produces a bound on the MSE of a $\mathcal{T}$-unbiased estimator, obtained by calculating the trace of (40) in the case $\boldsymbol{B}\boldsymbol{U} = \boldsymbol{0}$. This bound is given by

$$E\{\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|_2^2\} \geq \sigma^2 \operatorname{Tr}((\boldsymbol{H}_{\boldsymbol{\alpha}_0}^T \boldsymbol{H}_{\boldsymbol{\alpha}_0})^{-1}), \quad \|\boldsymbol{\alpha}_0\|_0 = s. \tag{44}$$

The most striking feature of (44) is that it is identical to the oracle MSE (14). However, the CRB is of additional importance because of the fact that the ML estimator achieves the CRB in the limit when a large number of independent measurements are available, a situation which is equivalent in our setting to the limit $\sigma \to 0$. In other words, an MSE of (44) is achieved at high SNR by the ML approach (8), as we will illustrate numerically in Section V. While the ML approach is computationally intractable in the sparse estimation setting, it is still implementable in principle, as opposed to $\widehat{\boldsymbol{\alpha}}_{\text{oracle}}$, which relies on unavailable information (namely, the support set of $\boldsymbol{\alpha}_0$). Thus, Theorem 1 gives an alternative interpretation to comparisons of estimator performance with the oracle.

Observe that the bound (44) depends on the value of $\boldsymbol{\alpha}_0$ (through its support set, which defines $\boldsymbol{H}_{\boldsymbol{\alpha}_0}$). This implies that some values of $\boldsymbol{\alpha}_0$ are more difficult to estimate than others. For example, suppose the ℓ_2 norms of some of the columns of $\boldsymbol{H}$ are significantly larger than the remaining columns. Measurements of a parameter $\boldsymbol{\alpha}_0$ whose support corresponds to the large-norm columns of $\boldsymbol{H}$ will then have a much higher SNR than measurements of a parameter corresponding to small-norm columns, and this will clearly affect the accuracy with which $\boldsymbol{\alpha}_0$ can be estimated. To analyze the behavior beyond this effect, it is common to consider the situation in which the columns $\boldsymbol{h}_i$ of $\boldsymbol{H}$ are normalized so that $\|\boldsymbol{h}_i\|_2 = 1$. In this case, for sufficiently incoherent dictionaries, $\operatorname{Tr}((\boldsymbol{H}_{\boldsymbol{\alpha}_0}^T \boldsymbol{H}_{\boldsymbol{\alpha}_0})^{-1})$ is bounded above and below by a small constant times s, so that the CRB is similar for all values of $\boldsymbol{\alpha}_0$. To see this, let μ be the coherence of $\boldsymbol{H}$ [1], defined (for $\boldsymbol{H}$ having normalized columns) as

$$\mu \triangleq \max_{i \neq j} \left| \boldsymbol{h}_i^T \boldsymbol{h}_j \right|. \tag{45}$$

By the Gershgorin disc theorem, the eigenvalues of $\boldsymbol{H}_{\boldsymbol{\alpha}_0}^T \boldsymbol{H}_{\boldsymbol{\alpha}_0}$ are in the range $[1 - s\mu, 1 + s\mu]$. It follows that the unbiased CRB (44) is bounded above and below by

$$\frac{s\sigma^2}{1+s\mu} \leq \sigma^2 \operatorname{Tr}((\boldsymbol{H}_{\boldsymbol{\alpha}_0}^T \boldsymbol{H}_{\boldsymbol{\alpha}_0})^{-1}) \leq \frac{s\sigma^2}{1-s\mu}. \tag{46}$$

Thus, when s is somewhat smaller than $1/\mu$, the CRB is roughly equal to $s\sigma^2$ for all values of $\boldsymbol{\alpha}_0$. As we have seen in Section II-B, for sufficiently small s, the worst-case MSE of practical estimators, such as BPDN and the DS, is $O(s\sigma^2 \log p)$. Thus, practical estimators come almost within a constant of the unbiased CRB, implying that they are close to optimal for all values of $\boldsymbol{\alpha}_0$, at least when compared with unbiased techniques.

B. Denoising and Deblurring

We next consider the problem (1), in which it is required to estimate not the sparse vector $\boldsymbol{\alpha}_0$ itself, but rather the vector $\boldsymbol{x}_0 = \boldsymbol{D}\boldsymbol{\alpha}_0$, where $\boldsymbol{D}$ is a known dictionary matrix. Thus, $\boldsymbol{x}_0$ belongs to the set $\mathcal{S}$ of (2). We assume for concreteness that $\boldsymbol{D}$ has full row rank and that $\boldsymbol{A}$ has full column rank. This setting

encompasses the denoising and deblurring problems described in Section II-A, with the former arising when $\boldsymbol{A} = \boldsymbol{I}$ and the latter obtained when $\boldsymbol{A}$ represents a blurring kernel. Similar calculations can be carried out when $\boldsymbol{A}$ is rank-deficient, a situation which occurs, for example, in some interpolation problems.

Recall from Section II-A the assumption that every $\boldsymbol{x} \in \mathcal{S}$ has a *unique* representation $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}$ for which $\boldsymbol{\alpha}$ is in the set $\mathcal{T}$ of (5). We denote by $\boldsymbol{r}(\cdot)$ the mapping from $\mathcal{S}$ to $\mathcal{T}$ which returns this representation. In other words, $\boldsymbol{r}(\boldsymbol{x})$ is the unique vector in $\mathcal{T}$ for which

$$\boldsymbol{x} = \boldsymbol{D}\boldsymbol{r}(\boldsymbol{x}) \quad \text{and} \quad \|\boldsymbol{r}(\boldsymbol{x})\|_0 \leq s. \tag{47}$$

Note that while the mapping $\boldsymbol{r}$ is well-defined, actually calculating the value of $\boldsymbol{r}(\boldsymbol{x})$ for a given vector $\boldsymbol{x}$ is, in general, NP-hard.

In the current setting, unlike the scenario of Section IV-A, it is always possible to construct an unbiased estimator. Indeed, even without imposing the constraint (2), there exists an unbiased estimator. This is the LS or maximum likelihood estimator, given by

$$\widehat{\boldsymbol{x}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}. \tag{48}$$

A standard calculation demonstrates that the covariance of $\widehat{\boldsymbol{x}}$ is

$$\sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}. \tag{49}$$

On the other hand, the FIM for the setting (1) is given by

$$\boldsymbol{J} = \frac{1}{\sigma^2}\boldsymbol{A}^T\boldsymbol{A}. \tag{50}$$

Since $\boldsymbol{A}$ has full row rank, the FIM is invertible. Consequently, it is seen from (49) and (50) that the LS approach achieves the CRB $\boldsymbol{J}^{-1}$ for unbiased estimators. This well-known property demonstrates that in the unconstrained setting, the LS technique is optimal among all unbiased estimators.

The LS estimator, like any unbiased approach, is also $\mathcal{S}$-unbiased. However, with the addition of the constraint $\boldsymbol{x}_0 \in \mathcal{S}$, one would expect to obtain improved performance. It is therefore of interest to obtain the CRB for the constrained setting. To this end, we first note that since $\boldsymbol{J}$ is invertible, we have $\mathcal{R}(\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{J}\boldsymbol{U}\boldsymbol{U}^T) = \mathcal{R}(\boldsymbol{U}\boldsymbol{U}^T)$ for any $\boldsymbol{U}$, and consequently (26) holds for any matrix $\boldsymbol{B}$. The bound (27) of Theorem 1 thus applies regardless of the bias gradient matrix.

For simplicity, in the following we derive the CRB for $\mathcal{S}$-unbiased estimators. A calculation for arbitrary $\mathcal{S}$-bias functions can be performed along similar lines. Consider first values $\boldsymbol{x} \in \mathcal{S}$ such that $\|\boldsymbol{r}(\boldsymbol{x})\|_0 < s$. Then, $\|\boldsymbol{r}(\boldsymbol{x}) + \delta\boldsymbol{e}_i\|_0 \leq s$ for any δ and for any $\boldsymbol{e}_i$. Therefore,

$$\boldsymbol{x} + \delta\boldsymbol{D}\boldsymbol{e}_i \in \mathcal{S} \tag{51}$$

for any δ and $\boldsymbol{e}_i$. In other words, the feasible directions include all columns of $\boldsymbol{D}$. Since it is assumed that $\boldsymbol{D}$ has full row rank, this implies that the feasible subspace $\mathcal{F}$ equals $\mathbb{R}^n$, and the matrix $\boldsymbol{U}$ of (20) can be chosen as $\boldsymbol{U} = \boldsymbol{I}$.

Next, consider values $\boldsymbol{x} \in \mathcal{S}$ for which $\|\boldsymbol{r}(\boldsymbol{x})\|_0 = s$. Then, for sufficiently small $\delta > 0$, we have $\|\boldsymbol{r}(\boldsymbol{x}) + \delta\boldsymbol{v}\|_0 \leq s$ if and only if $\boldsymbol{v} = \boldsymbol{e}_i$ for some $i \in \operatorname{supp}(\boldsymbol{r}(\boldsymbol{x}))$. Equivalently,

$$\boldsymbol{x} + \delta\boldsymbol{v} \in \mathcal{S} \text{ if and only if } \boldsymbol{v} = \boldsymbol{D}\boldsymbol{e}_i \text{ and } i \in \operatorname{supp}(\boldsymbol{r}(\boldsymbol{x})). \tag{52}$$

Consequently, the feasible direction subspace in this case corresponds to the column space of the matrix $\boldsymbol{D}_{\boldsymbol{x}}$ containing the s columns of $\boldsymbol{D}$ indexed by $\operatorname{supp}(\boldsymbol{r}(\boldsymbol{x}))$. From (7) we have $\operatorname{spark}(\boldsymbol{D}) > s$, and therefore the columns of $\boldsymbol{D}_{\boldsymbol{x}}$ are linearly independent. Thus the orthogonal projector onto $\mathcal{F}$ is given by

$$\boldsymbol{P} \triangleq \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{D}_{\boldsymbol{x}}(\boldsymbol{D}_{\boldsymbol{x}}^T\boldsymbol{D}_{\boldsymbol{x}})^{-1}\boldsymbol{D}_{\boldsymbol{x}}^T. \tag{53}$$

Combining these calculations with Theorem 1 yields the following result.

Theorem 3: Consider the estimation setting (1) with the constraint (2), and suppose $\operatorname{spark}(\boldsymbol{D}) > 2s$. Let $\widehat{\boldsymbol{x}}$ be a finite-variance $\mathcal{S}$-unbiased estimator. Then,

$$\begin{aligned} &\operatorname{Cov}(\widehat{\boldsymbol{x}}) \succeq \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1} && \text{when } \|\boldsymbol{r}(\boldsymbol{x})\|_0 < s, \\ &\operatorname{Cov}(\widehat{\boldsymbol{x}}) \succeq \sigma^2\left(\boldsymbol{P}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{P}\right)^{\dagger} && \text{when } \|\boldsymbol{r}(\boldsymbol{x})\|_0 = s. \end{aligned} \tag{54}$$

Here, $\boldsymbol{P}$ is given by (53), in which $\boldsymbol{D}_{\boldsymbol{x}}$ is the $n \times s$ matrix consisting of the columns of $\boldsymbol{D}$ participating in the (unique) s-element representation $\boldsymbol{D}\boldsymbol{\alpha}$ of $\boldsymbol{x}$.

As in Theorem 2, the bound exhibits a dichotomy between points having maximal and non-maximal support. In the former case, the CRB is equivalent to the bound obtained when the support set is known, whereas in the latter the bound is equivalent to an unconstrained CRB. This point is discussed further in Section VI.

V. Numerical Results

In this section, we demonstrate the use of the CRB for measuring the achievable MSE in the sparse estimation problem (4). To this end, a series of simulations was performed. In each simulation, a random 100×200 dictionary $\boldsymbol{H}$ was constructed from a zero-mean Gaussian IID distribution, whose columns $\boldsymbol{h}_i$ were normalized so that $\|\boldsymbol{h}_i\|_2 = 1$. A parameter $\boldsymbol{\alpha}_0$ was then selected by choosing a support uniformly at random and selecting the nonzero elements as Gaussian IID variables with mean 0 and variance 1. Noisy measurements $\boldsymbol{y}$ were obtained from (4), and $\boldsymbol{\alpha}_0$ was then estimated using BPDN (9), the DS (10), and the GDS (11). The regularization parameters were chosen as $\tau = 2\sigma\sqrt{\log p}$ and $\gamma = 4\sigma\sqrt{\log(p-s)}$, rules of thumb which are motivated by a theoretical analysis [11]. The MSE of each estimate was then calculated by repeating this process with different realizations of the random variables. The unbiased CRB was calculated using (44). In this case, the unbiased CRB equals the MSE of the oracle estimator (13), but as we will see below, interpreting (44) as a bound on unbiased estimators provides further insight into the estimation problem.

A first set of experiments was conducted to examine the CRB at various SNR levels. In this simulation, the ML estimator (8) was also computed, in order to verify its convergence to the CRB at high SNR. Since the ML approach is computationally prohibitive when p and s are large, this necessitated the selection of the rather low support size $s = 3$. The MSE and CRB were calculated for 15 SNR values by

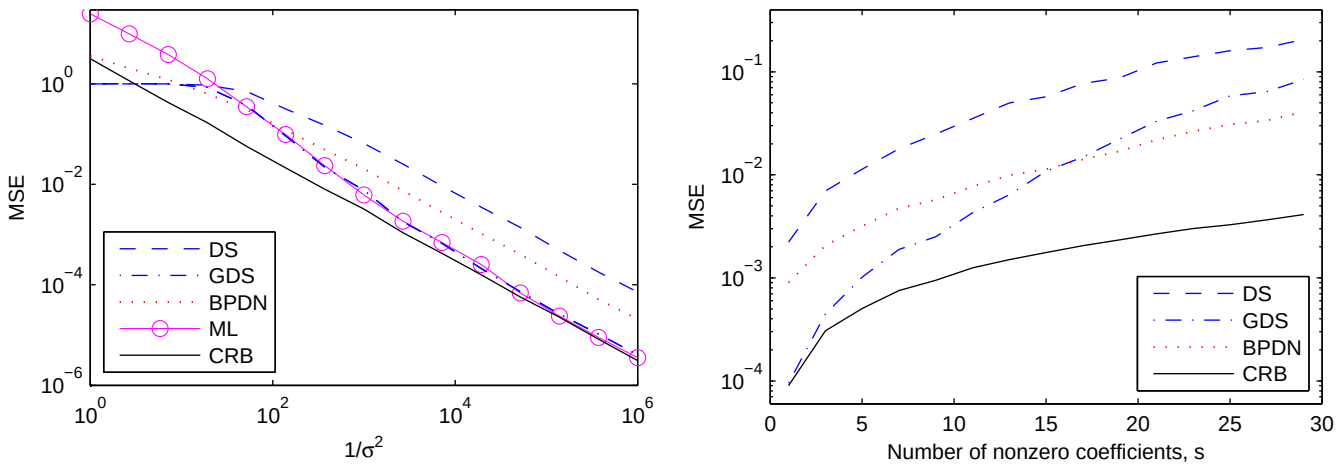

Fig. 2. MSE of various estimators compared with the unbiased CRB (44), for (a) varying SNR and (b) varying sparsity levels.

changing the noise standard deviation σ between 1 and 10^{-3}. The MSE of the ML approach, as well as the other estimators of Section II-B, is compared with the CRB in Fig. 2(a). The convergence of the ML estimator to the CRB is clearly visible in this figure. The performance of the GDS is also impressive, being as good or better than the ML approach. Apparently, at high SNR, the DS tends to correctly recover the true support set, in which case GDS (11) equals the oracle (13). Perhaps surprisingly, applying a LS estimate on the support set obtained by BPDN (which could be called a "Gauss–BPDN" strategy) does not work well at all, and in fact results in higher MSE than a direct application of BPDN. (The results for the Gauss–BPDN method are not plotted in Fig. 2.)

Note that some estimation techniques outperform the oracle MSE (or CRB) at low SNR. It may appear surprising that a practical technique such as the DS outperforms the oracle. The explanation for this stems from the fact that the CRB (44) is a lower bound on the MSE of *unbiased* estimators. The bias of most estimators tends to be negligible in low-noise settings, but often increases with the noise variance σ^2. Indeed, when σ^2 is as large as $\|\boldsymbol{\alpha}_0\|_2^2$, the measurements carry very little useful information about $\boldsymbol{\alpha}_0$, and an estimator can improve performance by shrinkage. Such a strategy, while clearly biased, yields lower MSE than a naive reliance on the noisy measurements. This is indeed the behavior of the DS and BPDN, since for large σ^2, the ℓ_1 regularization becomes the dominant term, resulting in heavy shrinkage. Consequently, it is to be expected that such techniques will outperform even the best unbiased estimator at low SNR, as indeed occurs in Fig. 2(a).

The performance of the estimators of Section II-B, excluding the ML method, was also compared for varying sparsity levels. To this end, the simulation was repeated for 15 support sizes in the range $1 \leq s \leq 30$, with a constant noise standard deviation of $\sigma = 0.01$. The results are plotted in Fig. 2(b). While a substantial gap exists between the CRB and the MSE of the practical estimators in this case, the general trend in both cases describes a similar rate of increase as s grows. Interestingly, a drawback of the GDS approach is visible in this setting: as s increases, correct support recovery becomes more difficult, and shrinkage becomes a valuable asset for reducing the sensitivity of the estimate to random measurement fluctuations. The LS approach practiced by the GDS, which does not perform shrinkage, leads to gradual performance deterioration.

Results similar to Fig. 2 were obtained for a variety of related estimation scenarios, including several deterministic, rather than random, dictionaries $\boldsymbol{H}$.

VI. Discussion

In this paper, we extended the CRB to constraint sets satisfying the local balance condition (Theorem 1). This enabled us to derive lower bounds on the achievable performance in various estimation problems (Theorems 2 and 3). In simple terms, Theorems 2 and 3 can be summarized as follows. The behavior of the CRB differs depending on whether or not the parameter has maximal support (i.e., $\|\boldsymbol{\alpha}\|_0 = s$). In the case of maximal support, the bound equals that which would be obtained if the sparsity pattern were known; this can be considered an "oracle bound". On the other hand, when $\|\boldsymbol{\alpha}\|_0 < s$, performance is identical to the unconstrained case, and the bound is substantially higher. We now discuss some practical implications of these conclusions. To simplify the discussion, we consider the case of unbiased estimators, though analogous conclusions can be drawn for any bias function.

When $\|\boldsymbol{\alpha}\|_0 = s$ and all nonzero elements of $\boldsymbol{\alpha}$ are considerably larger than the standard deviation of the noise, the support set can be recovered correctly with high probability (at least if computational considerations are ignored). Thus, in this case an estimator can mimic the behavior of the oracle, and the CRB is expected to be tight. Indeed, in the high SNR limit, the ML estimator achieves the unbiased CRB. On the other hand, when the support of $\boldsymbol{\alpha}$ is not maximal, the unbiasedness requirement demands sensitivity to changes in all components of $\boldsymbol{\alpha}$, and consequently the bound coincides with the unconstrained CRB. Thus, as claimed in Section III, in underdetermined cases no estimator is unbiased for all $\boldsymbol{\alpha} \in \mathcal{S}$.

An interesting observation can also be made concerning maximal-support points $\boldsymbol{\alpha}$ for which some of the nonzero elements are close to zero. The CRB in this "low-SNR" case corresponds to the oracle MSE, but as we will see, the bound is loose for such values of $\boldsymbol{\alpha}$. Intuitively, at low-SNR points, any attempt to recover the sparsity pattern will occasionally fail. Consequently, despite the optimistic CRB, it is unlikely that the oracle MSE can be achieved. Indeed, the covariance matrix of any finite-variance estimator is a continuous function of $\boldsymbol{\alpha}$ [22], and the fact that performance is bounded by the (much higher) unconstrained bound when $\|\boldsymbol{\alpha}\|_0 < s$ implies that performance must be similarly poor for low SNR.

This excessive optimism is a result of the local nature of the CRB: The bound is a function of the estimation setting only in an ε-neighborhood of the parameter itself. Indeed, the CRB depends on the constraint set only through the feasible directions, which were defined in Section III-B as those directions which do not violate the constraints for *sufficiently small* deviations. Thus, for the CRB, it is entirely irrelevant if some of the components of $\boldsymbol{\alpha}$ are close to zero, as long as $\mathrm{supp}(\boldsymbol{\alpha})$ is held constant.

A tighter bound for sparse estimation problems may be obtained using the Hammersley–Chapman–Robbins (HCR) approach [15], [28], [29], which depends on the constraints at points beyond the local neighborhood of $\boldsymbol{x}$. Such a bound is likely to yield tighter results for low SNR values, and will create a smooth transition between the regions of maximal and non-maximal support. However, the bound will depend on more complex properties of the estimation setting, such as the distance between $\boldsymbol{D\alpha}$ and feasible points with differing supports. The derivation of such a bound is a subject for further research.

ACKNOWLEDGEMENT

The authors would like to thank Yaniv Plan for helpful discussions. The authors are also grateful to the anonymous reviewers for their comments, which considerably improved the presentation of the paper.

REFERENCES

[1] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[2] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[3] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. LIX, pp. 1207–1223, 2006.

[4] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, 2007, with discussion.

[5] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[6] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[7] ——, "Image restoration by sparse 3D transform-domain collaborative filtering," in *Proc. SPIE Electronic Imaging '08*, no. 6812-07, San Jose, CA, Jan. 2008.

[8] M. Protter and M. Elad, "Image sequence denoising via sparse and redundant representations," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 27–36, Jan. 2009.

[9] M. Elad, J.-L. Starck, P. Querre, and D. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)," *J. Applied and Computational Harmonic Analysis*, vol. 19, pp. 340–358, Nov. 2005.

[10] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1998.

[11] Z. Ben-Haim, Y. C. Eldar, and M. Elad, "Near-oracle performance of basis pursuit under random noise," Mar. 2009. [Online]. Available: http://arxiv.org/abs/0903.4579

[12] E. J. Candès, "Modern statistical estimation via oracle inequalities," *Acta Numerica*, pp. 1–69, 2006.

[13] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[14] J. Shao, *Mathematical Statistics*, 2nd ed. New York: Springer, 2003.

[15] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Trans. Inf. Theory*, vol. 26, no. 6, pp. 1285–1301, Nov. 1990.

[16] T. L. Marzetta, "A simple derivation of the constrained multiple parameter Cramér–Rao bound," *IEEE Trans. Signal Process.*, vol. 41, no. 6, pp. 2247–2249, Jun. 1993.

[17] P. Stoica and B. C. Ng, "On the Cramér–Rao bound under parametric constraints," *IEEE Signal Process. Lett.*, vol. 5, no. 7, pp. 177–179, 1998.

[18] Z. Ben-Haim and Y. C. Eldar, "On the constrained Cramér-Rao bound with a singular Fisher information matrix," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 453–456, Jun. 2009.

[19] ——, "Performance bounds for sparse estimation with random noise," in *Proc. IEEE Workshop on Statistical Signal Processing*, Cardiff, Wales, UK, Sep. 2009.

[20] B. Babadi, N. Kalouptsidis, and V. Tarokh, "Asymptotic achievability of the Cramér–Rao bound for noisy compressive sampling," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 1233–1236, 2009.

[21] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[22] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer, 1998.

[23] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2202, March 4, 2003.

[24] Z. Ben-Haim and Y. C. Eldar, "Blind minimax estimation," *IEEE Trans. Inf. Theory*, vol. 53, no. 9, pp. 3145–3157, Sep. 2007.

[25] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inf. Theory*, to appear. [Online]. Available: http://arxiv.org/pdf/0807.4581

[26] Y. C. Eldar, "Compressed sensing of analog signals in shift-invariant spaces," *IEEE Trans. Signal Process.*, 2009, to appear. [Online]. Available: http://arxiv.org/abs/0806.3332

[27] K. Gedalyahu and Y. C. Eldar, "Low rate sampling schemes for time delay estimation," *IEEE Trans. Signal Process.*, May 2009, submitted. [Online]. Available: http://arxiv.org/abs/0905.2429

[28] J. M. Hammersley, "On estimating restricted parameters," *J. Roy. Statist. Soc. B*, vol. 12, no. 2, pp. 192–240, 1950.

[29] D. G. Chapman and H. Robbins, "Minimum variance estimation without regularity assumptions," *Ann. Math. Statist.*, vol. 22, no. 4, pp. 581–586, Dec. 1951.